Published online 6 April 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W59–W64***
doi:10.1093/nar/gks310

SAVoR: a server for sequencing annotation and visualization of RNA structures

Fan Li[1,2,3], Paul Ryvkin[1,4], Daniel M. Childress[4], Otto Valladares[4], Brian D. Gregory[1,2,3,*] and Li-San Wang[1,2,4,5,6,*]

[1]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, [2]Penn Genome Frontiers Institute, Perelman School of Medicine, University of Pennsylvania, [3]Department of Biology, [4]Department of Pathology and Laboratory Medicine, [5]Institute on Aging, Perelman School of Medicine, University of Pennsylvania and [6]Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Received January 19, 2012; Revised March 21, 2012; Accepted March 24, 2012

ABSTRACT

RNA secondary structure is required for the proper regulation of the cellular transcriptome. This is because the functionality, processing, localization and stability of RNAs are all dependent on the folding of these molecules into intricate structures through specific base pairing interactions encoded in their primary nucleotide sequences. Thus, as the number of RNA sequencing (RNA-seq) data sets and the variety of protocols for this technology grow rapidly, it is becoming increasingly pertinent to develop tools that can analyze and visualize this sequence data in the context of RNA secondary structure. Here, we present Sequencing Annotation and Visualization of RNA structures (SAVoR), a web server, which seamlessly links RNA structure predictions with sequencing data and genomic annotations to produce highly informative and annotated models of RNA secondary structure. SAVoR accepts read alignment data from RNA-seq experiments and computes a series of per-base values such as read abundance and sequence variant frequency. These values can then be visualized on a customizable secondary structure model. SAVoR is freely available at http://tesla.pcbi.upenn.edu/savor.

INTRODUCTION

The secondary structure of an RNA molecule comprises specific base pairing interactions encoded within the primary nucleotide sequence. The formation of secondary structure is vital to the maturation and function of many classes of RNAs. For example, the classic clover-leaf folding pattern of tRNAs is necessary for their function in translation, while the processing of multiple classes of small regulatory RNAs requires formation of specific secondary structures (1,2). Recently, the advent of high-throughput RNA sequencing (RNA-seq) has enabled unbiased, genome-wide studies of many RNA populations within the cell. RNA-seq and its variant protocols have been recently used to study a wide range of biological phenomena, including RNA silencing (3,4), RNA–protein interactions (5,6) and protein translation (7), to name a few. These experiments, along with several recent studies of RNA base pairing (4,8–10), have highlighted the functional significance of RNA structure on a global scale.

Although the importance of RNA secondary structure is clear, most existing tools for RNA-seq analysis, such as DESeq (11), Myrna (12), Cufflinks (13), and Galaxy (14), primarily report RNA-seq analyses in the context of linear transcript models and do not support a structure-based interpretation. On the other hand, tools that do enable visualization and annotation of RNA structure models [e.g. RNAstructure (15), RNAfold (16), etc.] are focused on the problem of RNA secondary structure prediction and are not easily applicable to analysis of RNA-seq data. To address this gap, we have developed Sequencing Annotation and Visualization of RNA structures (SAVoR), which neatly integrates common RNA-seq analyses with a structure-based annotation and visualization framework (Figure 1). To do this, SAVoR extracts sequencing data from user-specified RNA-seq alignment files and computes a series of per-nucleotide values such as read abundance and sequence variant frequency, which are then directly plotted on a customizable structural model. The entire process is streamlined via a simple web

*To whom correspondence should be addressed. Tel: +1 215 746 4398; Fax: +1 215 898 8780; Email: bdgregor@sas.upenn.edu
Correspondence may also be addressed to Li-San Wang. Tel: +1 215 746 7015; Fax: +1 215 573 3111; Email: lswang@mail.med.upenn.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

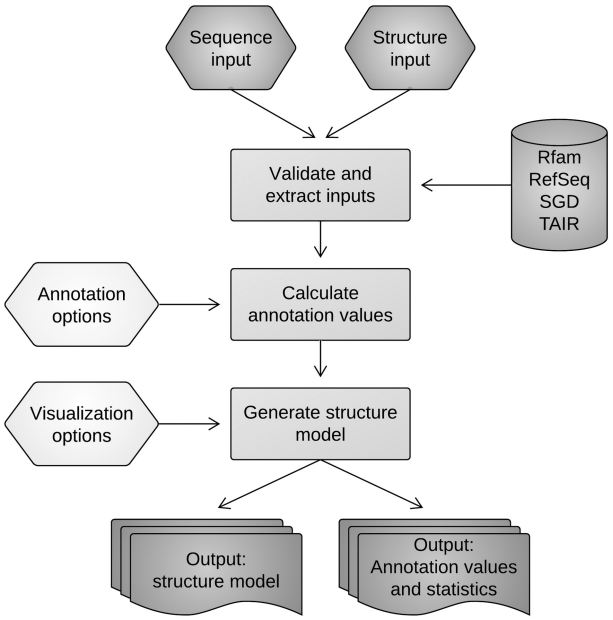

Figure 1. The SAVoR workflow. Upon validation of user input, the primary sequence and genomic location of the user-submitted transcript(s) are determined, and intersecting sequence reads are converted to the desired annotation values. The secondary structure is then determined and plotted with the specified visualization options.

interface and is completely platform independent. The uses of SAVoR range from a quick look at a transcript of interest to fully customized and annotated publication quality structure models.

SAVOR WEB SERVER

Input

SAVoR requires the user to enter an RNA transcript sequence and a secondary structure as input. The sequence can be entered as a primary sequence in plain text or FASTA format, an Rfam (17) or transcript [Refseq (18), SGD (19) or TAIR (20)] accession number, or the genomic location by chromosomal range and strand information (Figure 2). Currently, the input sequence is restricted to 20 000 nt in length. If an Rfam accession number is entered and multiple matching entries are located (often the case for repetitive RNA elements such as tRNAs), then SAVoR lists all matching entries from which the user can then select the desired locus. If a primary sequence is entered and any genome-based annotation is selected, then BLASTN (21) with '-gapopen 999 gapextend 999' and otherwise default parameters will be used to determine the genomic location of the input sequence. The user will be prompted to select from a list of the top 20 BLASTN results that pass an E-value cutoff of $1e^{-3}$. The user can use a simple drop-down menu to select the reference genome, which is used by SAVoR to retrieve database entries and primary sequence data. SAVoR currently supports the latest reference genome releases for human, mouse, *Drosophila melanogaster* (fruit fly), *Saccharomyces cerevisiae* (budding yeast), *Arabidopsis thaliana* and *Caenorhabditis elegans* and contains 3831 Rfam entries and 167 157 RefSeq/SGD/TAIR entries.

Specifying the secondary structure

Depending on the type of input sequence and RNA-seq data, the user has four options to specify how the model of RNA secondary structure is generated. For example, SAVoR can retrieve the secondary structure from the Rfam database when the input is an Rfam accession ID. Additionally, the RNAfold program can be used with or without experimental constraints to fold the sequence into its minimum free energy state. If the constrained option is selected, the $\log_2$ abundance ratios of structure-informative RNA-seq data sets (4,8–10) are used to derive experimental constraints for structure prediction. Specifically, in the resulting structure model, a base will be paired when the dsRNA-seq to ssRNA-seq abundance ratio for that nucleotide exceeds some given threshold and vice versa (4,9). SAVoR will then use RNAfold to find the best secondary structure model based on the given constraints. Finally, the user can enter a specific secondary structure using the common dot-parenthesis notation (22).

Generating per-base annotations

Next, the user specifies the type of annotation to be overlaid on the RNA secondary structure model. Importantly, SAVoR supports remote access to indexed BAM files, which are highly compact files that contain read alignments from an RNA-seq experiment. SAVoR directly extracts sequencing reads that intersect with the input RNA transcript without requiring the user to upload the entire BAM file. Extracted reads are then converted to per-nucleotide annotation values. SAVoR can generate four different annotation types based on BAM files from RNA-seq or other types of high-throughput sequencing experiments: (i) read abundance (number of reads that cover each nucleotide base), (ii) endpoint abundance (number of reads whose 5′ or 3′ endpoint occurs at each nucleotide base), (iii) per-base mismatch frequency and (iv) per-base normalized $\log_2$ abundance ratio (for this analysis, the user is required to enter URL of two BAM files, which will be used by SAVoR to compute ratios).

It is worth noting that when $\log_2$ abundance or abundance ratio is selected, pseudo counts (adding 1 to the count of every position) are used to avoid numerical errors.

Alternatively, the user can upload a text file of custom annotation values using the UCSC Genome Browser BED format, a flexible tabular file format for genomic locations and associated data (http://genome.ucsc.edu/FAQ/FAQformat.html#format1). Currently, the BED-format file is limited to 5 Mb in size. Finally, the user can select from a series of visualization options that specify the markup and color scheme of the output structural model (these options are described in detail on the SAVoR website). Figure 2 shows an example input page that uses genomic coordinates for sequence input, a custom dot-parenthesis structure, and read coverage annotation with default visualization settings. The user can try out this sample input by clicking the 'Sample input' link on the SAVoR home page.

Figure 2. SAVoR can be used to visualize RNA-seq results in the context of predicted RNA secondary structures. In this example, the 'User Input' tab of the SAVoR web interface is populated with a sample input. The desired transcript is specified by UCSC-style genomic coordinates, and a custom structure is provided in dot-parenthesis notation. The 'coverage' annotation type has been selected and the URL for a web-accessible BAM file of read alignments provided. All optional visualization settings have been selected, and a blue-red color scheme will be used.

Output

After the user submits the input data, the 'Output' tab is automatically displayed (Figure 3). Progress indicators are shown for each step of the SAVoR workflow, along with warnings at steps that may require additional processing time. SAVoR validates all user-supplied data (e.g. that the input sequence only contains valid nucleotide characters) and reports any detected anomalies to help the user fix any errors in the input. Upon completion, the output structural model is directly displayed in the web browser, along with the calculated annotation values in tabular format. A legend showing the color scheme and annotation type is displayed in the top left corner of the output model. The sequence and structure are displayed using the default layout by RNAplot (16) with annotation values overlaid. The 5′ and 3′ ends of the transcript, as well as the position of every 10th nucleotide, are marked to facilitate location of a specific region of interest. The entire model can be scaled and panned as desired using standard browser tools.

Links to the output structure model in SVG, PDF, and PNG formats, and the annotation values in plain text format are provided as well. Importantly, files generated by each user submission are uniquely named and can only be viewed via these output links. The results are kept by the server for at least 72 h. If changes to the input data are

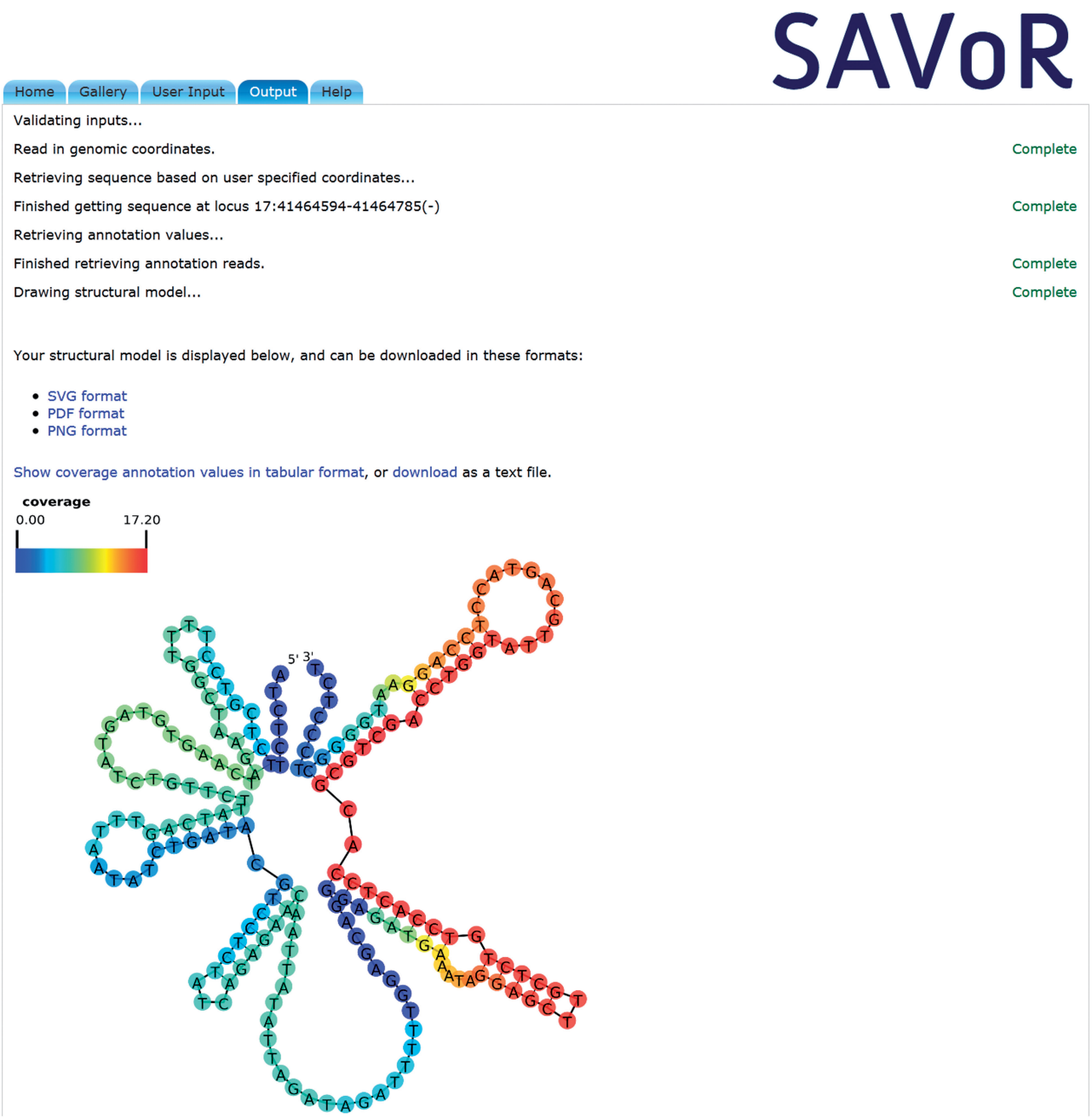

Figure 3. SAVoR produces highly informative, user specified, annotated models of RNA secondary structure. The 'Output' tab of the SAVoR web interface is dynamically rendered to show progress indicators and error messages. Each step of the pipeline is indicated, and a 'Complete' message appears upon successful completion. When the entire process is finished, the output (an annotated model of RNA secondary structure) is displayed, and links for downloading the output files are provided. Additionally, the calculated per-base annotation values are displayed in tabular form and can be downloaded as a plain text file.

desired, the user can simply click on the 'User Input' tab and directly modify the stored input values. While resubmission of the input form will result in rerunning of the entire SAVoR pipeline and generation of new output files, a typical SAVoR run requires <30 s for a 1 kb sequence.

Example uses

While we have streamlined the workflow design to strengthen its accessibility, SAVoR is also very flexible. We describe three example use cases to illustrate this point. Corresponding figures can be found in the Gallery page on the SAVoR website.

(i) Visualizing read distribution across a known transcript: The user specifies an Rfam or RefSeq ID, the 'coverage' annotation option, and read alignments as a BAM file. This type of model can be used to look for biases in read distribution such as those derived from small RNAs produced from a precursor transcript.

(ii) Comparing experimental and computational base pairing predictions: The user specifies the 'RNAfold' structure prediction method and 'log-ratio' annotation option, and provides two BAM files (containing structure-informative RNA-seq data) as input. The resulting output can be used to compare base pairing predictions from a free-energy based computational approach (RNAfold) with experimentally derived base pairing data (log-ratio).

(iii) Visualizing single nucleotide polymorphisms (SNPs): The user uploads a UCSC Genome Browser BED format file of customized per-nucleotide values along with any set of sequence and structure inputs. For example, we can upload a file-containing SNP coordinates and use this to color known SNPs on the secondary structure; this allows the user to examine if population diversity correlates with predicted or experimentally determined RNA structural constraints.

Implementation

The SAVoR web server runs Apache 2.2.3 on a CentOS 5.7 machine with 2× Intel Xeon E5450 3.00 GHz processors and 16 GB RAM. Asynchronous JavaScript and XML (AJAX) technology is used to dynamically render PHP output into formatted HTML. A local MySQL database is used to store Rfam and Refseq/SGD/TAIR entries, and a local installation of BLAST+ is used to retrieve sequence and genomic locus information. Structure prediction is optionally performed using a local installation of RNAfold (version 1.8.4), and backbone layout is done using RNAplot. SAMtools (23) is used to extract annotation values from BAM files, and custom Perl and Ruby scripts are used to process BED files. Inkscape (version 0.47) is used to convert from the native SVG format to publication quality PDF and PNG output files. SAVoR has been tested extensively, and the internal programs were used to generate annotated models of RNA secondary structure in our recent publications (4,9).

CONCLUSIONS

The incredible power and versatility of high-throughput RNA-seq approaches have spurred many insights into RNA function, biogenesis, and structure, and offer almost endless possibilities for future studies of RNA biology (24–27). Interpretation of these data is fast becoming a bottleneck, and substantial efforts to aid in this process are currently necessary. With SAVoR, we have developed a unique and user-friendly interface to streamline the interpretation of RNA-seq data in the context of RNA secondary structure. Specifically, our web server directly computes per-nucleotide quantities from RNA-seq data sets and overlays these annotation values on a structural model. The uses of this web-based tool range from quick checks of data quality to production of fully customized publication quality figures, and will aid researchers in many aspects of RNA-seq analysis. We plan to extend SAVoR to directly retrieve annotations from the UCSC Genome Browser and other public genomic databases, thereby removing the need for users to generate their own annotation files and improving accessibility to different data sources beyond sequencing alignments. We also plan to add other methods for structure prediction and visualization such as conservation-informed prediction (28,29), and implement other RNA secondary structure layout options including circular and linear structure plots. These additional features will further aid interpretation of genome-scale data in the context of RNA secondary structure.

ACKNOWLEDGEMENTS

We thank members of Gregory and Wang labs, Mingyao Li, John Hogenesch, Chris Stoeckert, and other participants of the HTS working group at Penn for their constructive comments and suggestions.

FUNDING

Funding for open access charge: Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine. This work was supported by National Science Foundation [MCB-1053846 to B.D.G.]; Penn Genome Frontiers Institute [pilot award to B.D.G. and L.S.W.]; National Institutes of Health [NIA AG10124 to B.D.G.and L.S.W.; NHGRI 5T32HG000046-13 to F.L. and P.R.]; SmithKline Beecham Center of Excellence in Geriatric Medicine through the Penn Institute on Aging [to M.C.].

Conflict of interest statement. None declared.

REFERENCES

1. Cruz,J.A. and Westhof,E. (2009) The dynamic landscapes of RNA architecture. *Cell*, **136**, 604–609.
2. Sharp,P.A. (2009) The centrality of RNA. *Cell*, **136**, 577–580.
3. Bracken,C.P., Szubert,J.M., Mercer,T.R., Dinger,M.E., Thomson,D.W., Mattick,J.S., Michael,M.Z. and Goodall,G.J. (2011) Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage. *Nucleic Acids Res.*, **39**, 5658–5668.
4. Zheng,Q., Ryvkin,P., Li,F., Dragomir,I., Valladares,O., Yang,J., Cao,K., Wang,L.S. and Gregory,B.D. (2010) Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.*, **6**.
5. Lebedeva,S., Jens,M., Theil,K., Schwanhäusser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
6. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M. Jr, Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
7. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
8. Kertesz,M., Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.
9. Li,F., Zheng,Q., Ryvkin,P., Dragomir,I., Desai,Y., Aiyer,S., Valladares,O., Yang,J., Bambina,S., Sabin,L.R. *et al.* (2012)

Global analysis of RNA secondary structure in two metazoans. *Cell Rep.*, **1**, 69–82.
10. Underwood,J.G., Uzilov,A.V., Katzman,S., Onodera,C.S., Mainzer,J.E., Mathews,D.H., Lowe,T.M., Salama,S.R. and Haussler,D. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.
11. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
12. Langmead,B., Hansen,K.D. and Leek,J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, **11**, R83.
13. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
14. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
15. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
16. Hofacker,I.L. and Stadler,P.F. (2006) Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, **22**, 1172–1176.
17. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
18. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
19. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
20. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
21. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
22. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, **125**, 167–188.
23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
24. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
25. Westhof,E. and Romby,P. (2010) The RNA structurome: high-throughput probing. *Nat. Methods*, **7**, 965–967.
26. Schmitz,R.J. and Zhang,X. (2011) High-throughput approaches for plant epigenomic studies. *Curr. Opin. Plant Biol.*, **14**, 130–136.
27. Saxena,A. and Carninci,P. (2011) Whole transcriptome analysis: what are we still missing? *Wiley Interdiscip* Rev. *Syst. Biol. Med.*, **3**, 527–543.
28. Engelen,S. and Tahi,F. (2010) Tfold: efficient in silico prediction of non-coding RNA secondary structures. *Nucleic Acids Res.*, **38**, 2453–2466.
29. Gruber,A.R., Findeiβ,S., Washietl,S., Hofacker,I.L. and Stadler,P.F. (2010) RNAz 2.0: Improved noncoding RNA detection. *Pac. Symp. Biocomput.*, **15**, 69–79.